An Approximation to Voice Aperiodicity

OSAMU FUJIMURA, MEMBER, IEEE

Abstract—Aperiodicity in voiced segments of speech may be ascribed to different causes. The magnitude of pitch perturbation is different in different spectral ranges of the signal. To see whether pitch perturbation can be effectively simulated by partially replacing voiced excitation by random noise, in appropriate frequency-time portions, experimental tests have been made on a computer-simulated channel vocoder. The buzz-hiss decision was made separately for three different frequency portions of the signal. The cepstrum technique was used for pitch detection, and separate buzz-hiss switching decisions were made at the synthesizer for each frequency portion. The switching thresholds were controlled, and deliberately "devoiced" versions were compared with regular vocoded speech. The fundamental frequency was determined by the lowband cepstrum. The result shows that partial devoicing of the high-frequency ranges definitely improves speech quality. Further, a comparatively large amount of devoicing is perceptually tolerable.

Manuscript received September 27, 1967. This paper was presented at the 1967 Conference on Speech Communication and Processing, Cambridge, Mass. This work was carried out at Bell Telephone Laboratories, Inc., Murray Hill, N. J., where the author was a consultant.

The author is with the Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, University of Tokyo, Hongo, Tokyo, Japan.

SEVERAL FACTORS may cause perturbation of the periodicity of speech waveforms in the so-called "voiced segments." The speech waveforms in these segments are produced by use of the vocal-cord vibration as the sound source. The vocal-cord vibration is never perfectly periodic, and the degree of aperiodicity, which depends to a large extent on the individual talker and conditions of phonation, contributes significantly to the voice quality, naturalness, and talker characteristics.[1] In speech band-compression systems, appropriate signal processing concerning this factor seems essential in retaining high quality of the reproduced speech signal.

In recent studies, it has been noted that the extent of aperiodicity may be regarded as frequency dependent.[2] It varies with time, not only because of changes in laryngeal conditions and those in vocal-tract conditions as independent factors, but also due to their mutual interaction. Thus, the actual aperiodicity in various degrees and of different natures is distributed in the time-frequency domain in a complex manner. In particular, there are "devoiced patches" in the two-dimensional space.[3]

Devoicing of speech segments is not necessarily limited to phonological devoicing where a language-dependent rule predicts the phonetic voiceless features, such as the diffuse vowels surrounded by voiceless phonemes in Japanese. There are cases of partial devoicing that are predictable by physical (phonetic) reasons.[4] Thus, not only the voiced fricative [z] but also the tense front vowel [i] tend to produce turbulent noise that has more energy than higher harmonics of the voice signal in a frequency range, say, above 3000 Hz. The sound, in such cases, can be approximated in that frequency domain simply by noise.

Close inspection of frequency analyses by a narrow-bandpass filter reveals that in speech samples in general, there are many segments for which a portion of the higher frequency range (e.g., above 1000 Hz) is almost completely devoiced (i.e., aperiodic), whereas it is hardly encountered that periodicity is found only in a higher frequency region without accompanying a voiced baseband signal. Sometimes, however, a devoiced patch can be found in an intermediate-frequency region (e.g., 1–2 kHz) rather than in higher frequencies. An example is shown in Fig. 1.

[1] O. Fujimura, "The Nagoya group of research on speech communication, a review of some of their publications," *Phonetica (Basel, Switzerland)*, vol. 7, no. 2/3, pp. 160–162, 1961.

[2] ——, "Speech coding and the excitation signal," *1966 IEEE Internat'l Commun. Conf., Digest of Technical Papers*, p. 49, 1966.

[3] It may be noted, incidentally, that a vertical striation in the wideband sound spectrogram does not necessarily prove true periodicity, because a random noise produced at the constriction may be simply amplitude-modulated in synchronism with the glottal movements. This kind of situation has been found in some actual speech samples.

[4] E. Fischer-Jørgensen, "Beobachtungen über den Zusammenhang zwischen Stimmhaftigkeit und intraoralem Luftdruck Zeitschrift für Phonetik," *Sprachwissenschaft und Kommunikationsforschung* (Basel, Switzerland), vol. 16, no. 1–3, pp. 19–36, 1963.

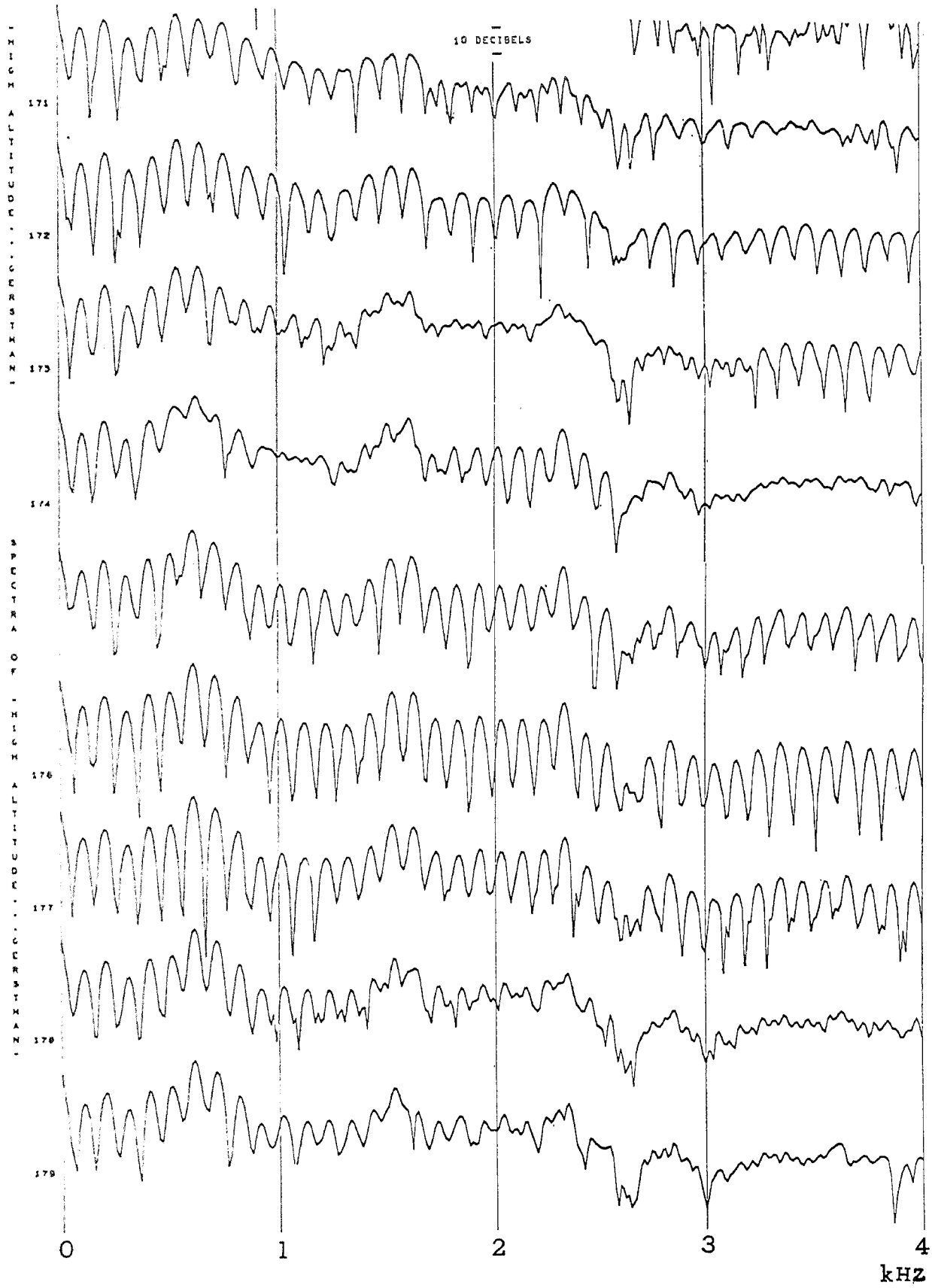

Fig. 1. A series of log-spectra representing a vowel segment [æ]. Time runs from top to bottom.

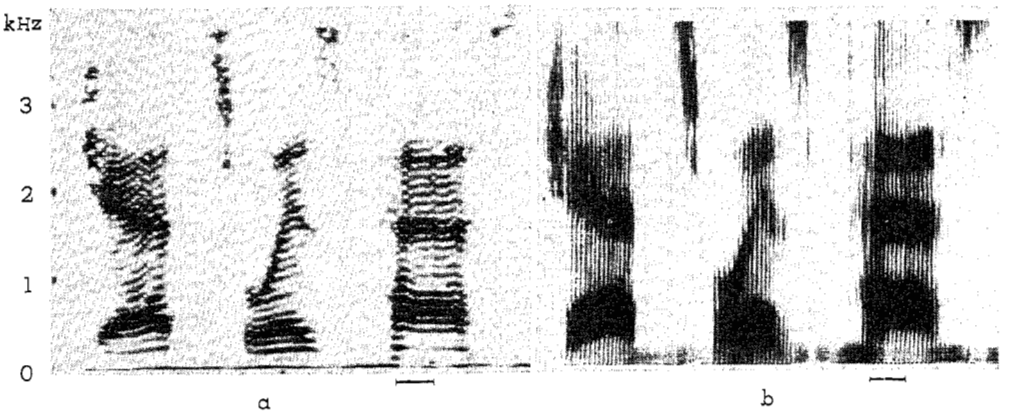

Fig. 2. Sound spectrograms for a portion (in italic) of the sentence "High altitude *jets whizz past* screaming," (a) by use of a narrowband (45-Hz) filter, and (b) by a wideband (300-Hz) filter.

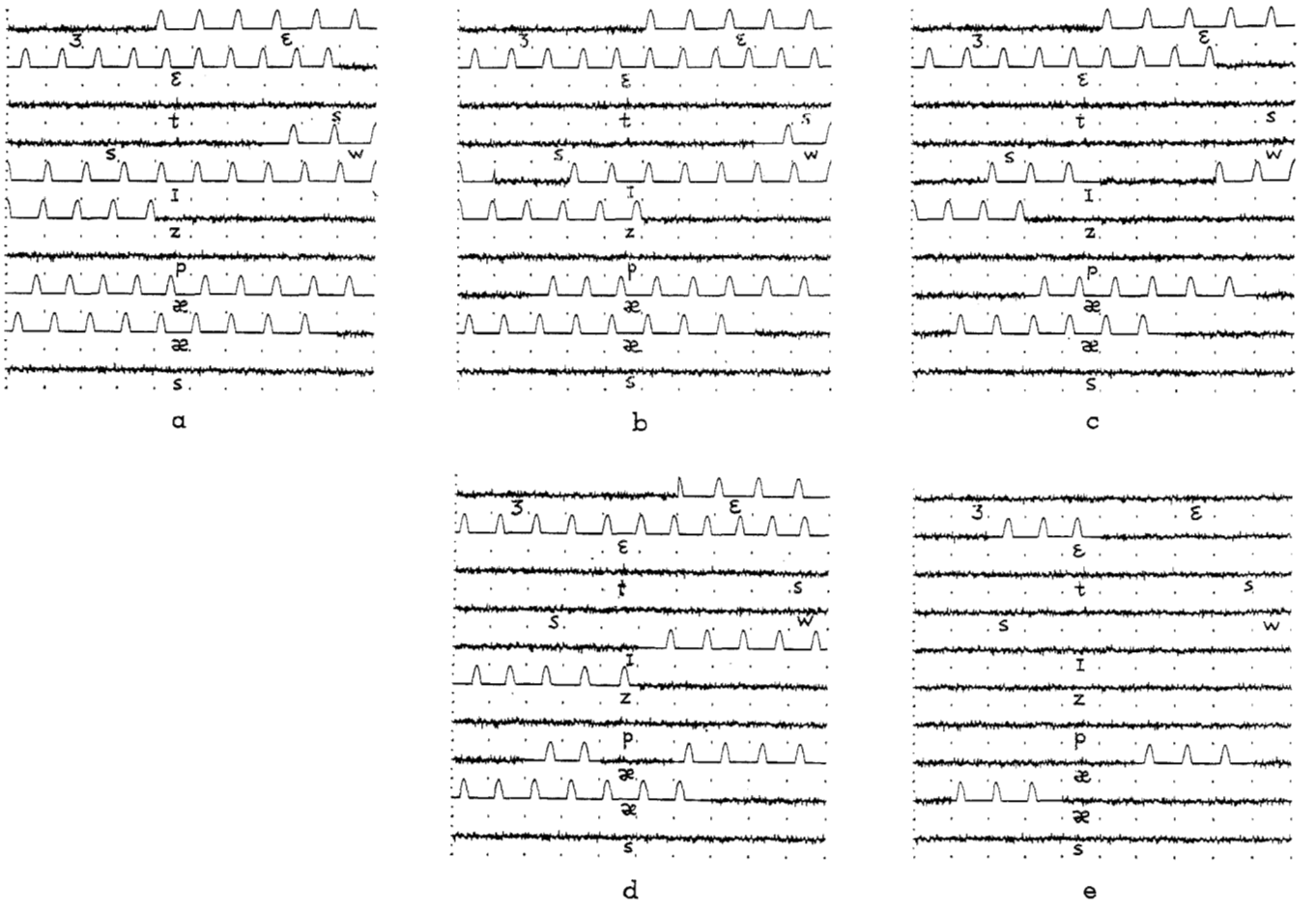

Fig. 3. Buzz–hiss combined source signals for the baseband (a), midband (b) and (d), and highband (c) and (e). The set (a), (b), and (c) is for an optimal extent of devoicing, whereas the set (a), (d), and (e) was used for an excessively devoiced version.

This analysis has been obtained by a digital computation, and the series of spectra are given in terms of the logarithm of the component amplitudes. The sample is a segment (identified by a horizontal bar underneath the spectrograms shown in Fig. 2) corresponding to a portion of an open vowel /æ/ in the word "past" of the sentence, "High altitude jets whizz past screaming," and represents a typical utterance of a normal male speaker of American English. It was rather surprising to find this partial devoicing here (see the third, fourth and the last two spectral samples in Fig. 1) in the form of disappearance of the harmonic structure. The speech sample sounds quite normal and convincingly well voiced; also, it is for a wide open articulation where no constriction along the vocal tract can be expected. The frequency portion in question is not weak in energy, and a regular periodic component could not be obscured under general noise. Sound spectrograms, as shown in Fig. 2, particularly one using a narrowband filter, however, reveals noticeable perturbations corresponding to the moments in question (near the left extreme and at the center of the voiced segment). Apparently, they are due to momentary irregularities of the vocal-cord movement, but the baseband and, in fact, in some cases higher frequency portions, are undoubtedly periodic. In most frequently encountered cases, however, the higher the frequency portion is, the more often the devoicing occurs.

It is questionable, however, if an exact specification of the frequency location of the devoiced patch, or even to some extent of its time location, is perceptually appreciable when it is above the baseband. It is anticipated, also, that a statistical approximation of the extent of devoicing, if given appropriately, may be effective in reproducing a natural voice quality. On the other hand, experience in vocoder experiments tells us that a large extent of deviation from the average devoicing, particularly artificially perfect voicing of all frequency components in "voiced" segments, tend to make the synthesized sound mechanically buzzy and unnatural. Vocoder engineers in designing a practical system often resort to a constant addition of some noise for the source signal. There may be, however, some crucial portions which either have to be, or must not be, devoiced. For this case, a computer-simulated channel vocoder was used for a preliminary experiment to test perceptual effect of a rough approximation of partial devoicing in various degrees.

A regular 14-channel vocoder was computer simulated and driven by a source signal that was produced according the cepstrum analysis of the speech sample.[5] One sample sentence uttered by a normal male talker, as mentioned earlier, was used for a close inspection. The log-spectrum and the cepstrum of the entire utterance were made and carefully examined. A threshold in terms of the height of cepstrum peak for the voiced–voiceless judgment was determined. A program for pitch determination and insertion of random noise in place of pitch pulses was run to produce a source signal which was based on a pitch-hiss dichotomy principle (see Fig. 3).

The same course of cepstrum analysis and voiced–voiceless decision was applied individually to three different frequency portions of the speech signal, viz., for the baseband approximately 250 to 1000 Hz, for the midband approximately 1000 to 2000 Hz, and for the highband approximately 2000 to 4000 Hz. This frequency selection was achieved by use of weighting functions as different windows through which the log-spectrum was viewed for the cepstrum computation. Some special techniques concerning the window and also cepstrum computation were employed to avoid undesired artifacts that might be caused by limiting the integration domain as described.

A source signal for vocoded samples was produced individually for each of the three frequency bands, and was applied for the proper band consisting of a subset of the 14 channels. The pulses of the source signals in the two higher bands were positioned, not by separate determinations of pitch values, but by adopting the pitch values determined by the baseband. Portions of the three source signals are illustrated in Fig. 3 for two different vocoding conditions. As the controlled condition, different sets of voiced–voiceless detection thresholds were used, and, consequently, we obtained versions of different extents of devoicing. In one of the vocoded versions, the threshold values were set at optimal values, in the sense that the decisions resulted in most reasonable runs of voiced and voiceless portions, when consulted with the natural sample by perceptual and spectrographic inspections. The source signals for this condition are illustrated in Fig. 3(a), (b), and (c) for the baseband, midband, and highband, respectively. In one of the versions that was made and examined, the voicing decisions were made identical for different bands and were all based on the baseband decisions. In one version, the highband was completely devoiced, and in another, both high- and midbands were completely devoiced. Some versions were more or less artificially devoiced only in mid- and highbands by making the thresholds of voicing judgments higher than the optimal values. One such case is shown in Fig. 3(d) and (e), the former illustrating the midband source and the latter, the highband source.

The resulting vocoded speech samples were listened to carefully by vocoder specialists in informal listening tests, and it was concluded that the optimal setting of the threshold values gave rise to a distinctly better voice quality than the sample produced by the single excitation signal which was derived from the basebar

[5] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293–309, 1967.

signal. Some of the excessively devoiced versions with threshold values too high, in particular, the version shown in Fig. 3(d) and (e) [the baseband source being the same as (a)], turned out to be quite acceptable and better than the single-source excitation. When the degree of artificial devoicing became too extreme, however, the degradation was noticed, but it was not too obvious even for the versions with the completely devoiced highband or mid- and highbands.

As the result of this preliminary study, we may conclude the following. The partial devoicing of various causes as observed in natural utterances has perceptual effects. A crude approximation of the aperiodicity can be made by distributing patches of random noise signals in the time-frequency space of the speech signal, while the other portions are made completely voiced (i.e., quasi-periodic). If the distribution of this dichotomous buzz–hiss selection is made appropriately, its application to vocoder techniques may be significantly effective in improving the resynthesized voice quality.

Acknowledgment

The author expresses his particular thanks to Dr. M. R. Schroeder, Dr. J. L. Flanagan, Dr. A. M. Noll, Dr. R. M. Golden, and Dr. M. M. Sondhi, all of Bell Telephone Laboratories, Inc., Murray Hill, N. J., for their effective cooperation in the experiment. He also acknowledges appreciation of efficient programming work carried out by Miss Sue Hanauer and Miss Lorinda Landgraf.

Osamu Fujimura (M'62) was born in Tokyo, Japan, on August 29, 1927. He received the B.S. degree in 1952, and the D.Sc. degree in 1962 from the University of Tokyo, Tokyo, Japan.

From 1952 to 1958 he was engaged as a Research Assistant in fundamental research in speech at the Kobayashi Institute of Physical Research. From 1958 to 1965 he was Assistant Professor at the University of Electro-Communications, Tokyo. At the same time, he was also appointed to the Division of Sponsored Research Staff, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge. He engaged in speech research at the Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, Sweden, as a Guest Researcher. He has also been a consultant in speech research at Bell Telephone Laboratories. In November, 1965, he was appointed to Professor in charge of the Speech Science Dept., Research Institute of Logopedics and Phoniatrics, University of Tokyo.

Dr. Fujimura is a fellow of the Acoustical Society of America.